THE EUROPEAN JOURNAL OF APPLIED ECONOMICS

EJAE 2024, 21(1): 110 - 120
ISSN 2406-2588
UDK: 330.133.2(73)"2003/2022"
330.43:519.237.8
DOI: 10.5937/EJAE21-45523
Original paper/Originalni naučni rad

CLUSTERING BASED ON THE ARCHETYPAL ANALYSIS

Beáta Stehlíková*

Faculty of Mathematics, Physics and Informatics,
Comenius University in Bratislava Mlynská dolina,
Bratislava, Slovakia

Abstract:

Archetypal analysis is a dimensionality reduction technique, which is based on finding a small number of representative elements, called archetypes. The observations are then approximated by convex combinations of the archetypes. The coefficients of the convex combinations can be therefore interpreted as probabilities of discrete random variables. The values of the variables identify the classes, represented by the archetypes, to which the observation belongs. Based on this interpretation, we propose to use the Hellinger distance between probability distributions to measure the distance between the observations in the dataset and to use it as an input to clustering. We apply this procedure to monthly data of zero-coupon yield curves in 2003-2022. We identify the archetypal yield curves and cluster the observed curves into six clusters. Since the observations are measured in time, the resulting clustering also gives a segmentation of the time period under consideration.

Article info:

Received: July 15, 2023
Correction: September 07, 2023
Accepted: November 03, 2023

Keywords:
archetypal analysis,
probability distribution,
Hellinger distance,
clustering,
yield curve

INTRODUCTION

In the data analysis setting proposed by Cutler and Breiman (1994), an *archetype* is a data point which is a typical example of a particular kind of observation based on its characteristics, *i.e.*, the observed variables or features. All observations in the dataset are then written as convex combinations of the archetypes.

Different modifications of the original archetypal analysis have been proposed, they include archetypoid analysis where the role of archetypes is given to selected observed data points (Vinué, Epifanio & Alemany, 2015), extensions to functional data (Epifanio, 2016), integer, binary and probability data (Seth & Eugster, 2016), ordinal data (Fernández, Epifanio and McMillan, 2021), weighted and robust archetypal analysis (Eugster & Leish, 2011), hierarchical archetypal analysis (Canhasi & Kononenko, 2016), dealing with missing data (Epifanio, Ibánez & Simó, 2020). Application of archetypal analysis include sports analytics (Vinué & Epifanio, 2017), analysis of publications in the field of economics (Seiler & Wohlrabe, 2013), weather and climate extremes (Hannachi & Trendafilov, 2017), brain responses (Tsanousa, Laskaris & Angelis, 2015), influenza outbreaks (Mokhtari, Landguth, Anderson & Stone, 2021), returns and volatility of S&P 500 stocks (Moliner & Epifanio, 2019).

*E-mail: stehlikova@fmph.uniba.sk

Our paper aims to apply the Hellinger distance, known from the comparison of probability distributions to the results of the archetypal analysis. Since the observations can be characterized by weights obtained as an output of the archetypal analysis, which has a character of discrete probability distributions, the Hellinger distance will make it possible to cluster the observations. We use this methodology to analyse the data about zero-coupon yield curves in the USA in 2003-2022 (*i.e.*, 20 years) with maturities ranging from 1 month to 30 years. Although the paper is meant primarily as an exploratory data analysis, an analysis of yield curves brings a natural question about its relation to well-known factors describing yield curves – level, slope and curvature (Litterman & Scheinkman, 1991). We expect the archetypes to have a certain relation to these factors and we will explore it on the dataset under study.

The paper is organized as follows. The methodology section reviews the archetypal analysis and Hellinger distance and then describes our proposed approach for their combination with illustrative examples. The application section describes the yielded data and provides the results of the archetypal analysis and the hierarchical clustering using the Hellinger distance. We show the yield curves for each of the clusters and present the segmentation of the time period considered.

METHODOLOGY

Archetypal analysis

Given *n* observations of *m* variables, the archetypes form a set of *p* elements (typically much smaller than the number of observations *n*), with each archetype being an *m*-dimensional vector (*i.e.*, it has its value for each variable of the dataset). For the given archetypes, the observations are approximated by convex combinations of the archetypes by minimizing the norm of the matrix of the differences between them. At the same time, the archetypes are constructed as convex combinations of the observations. Therefore, the optimization problem consists of finding two matrices of optimal weights: one defines the archetypes as convex combinations of data, and the latter approximates the observations by convex combinations of the archetypes. The optimal solution is not given by a closed-form expression but needs to be found using numerical methods. The original paper introducing the archetypal analysis (Cutler & Breiman, 1994) addressed this question, a modification, focusing on selecting the initial candidates for archetypes, has been suggested (Bauckhage & Thurau, 2009). We use the implementation from the *archetype* package (Eugster & Leisch, 2009) of R software with the default parameters, *i.e.*, the default value of the penalization in solving the convex least squares problem and using the QR decomposition when solving the system of linear equations.

There is no universal rule for selecting the number of archetypes. We use the elbow criterion, common in statistical analyses and suggested also in the presentation of the R package which we use (Eugster & Leisch, 2009): The algorithm is run for several possible values and the residual sum of squares is plotted against the number of archetypes. The final number of archetypes is selected by determining, where the curve starts to decrease more slowly. Additionally, since the initial archetypes are selected randomly, we run the algorithm several times for each number of archetypes under consideration. Furthermore, this helps to avoid the local minima in the optimization process (Eugster & Leisch, 2009).

Hellinger distance between probability distributions

In general, a statistical distance is a quantification of a "distance" (from a chosen point of view) between two probability distributions. Many statistical distances are not metrics, because they are not symmetric, typical examples are divergences such as Kullback–Leibler divergence (Kullback & Leibler, 1951). They give the distance of a given distribution from one particular reference distribution. However, in applications without a reference distribution, this is not a desirable property, and it is preferred to use a statistical distance which is a metric. An example of such a statistical distance is Hellinger distance (Hellinger, 1909), which we use for our analysis. It has been successfully used in various applications in existing literature, such as approximation of correlation matrices (Ning *et al*., 2013), centrality measures in social networks (Taheri *et al*., 2019), distinguishing equilibrium and disordered states in time series (Vamvakaris *et al*., 2018) or multicriteria decision making (Lourenzutti & Krohling, 2014).

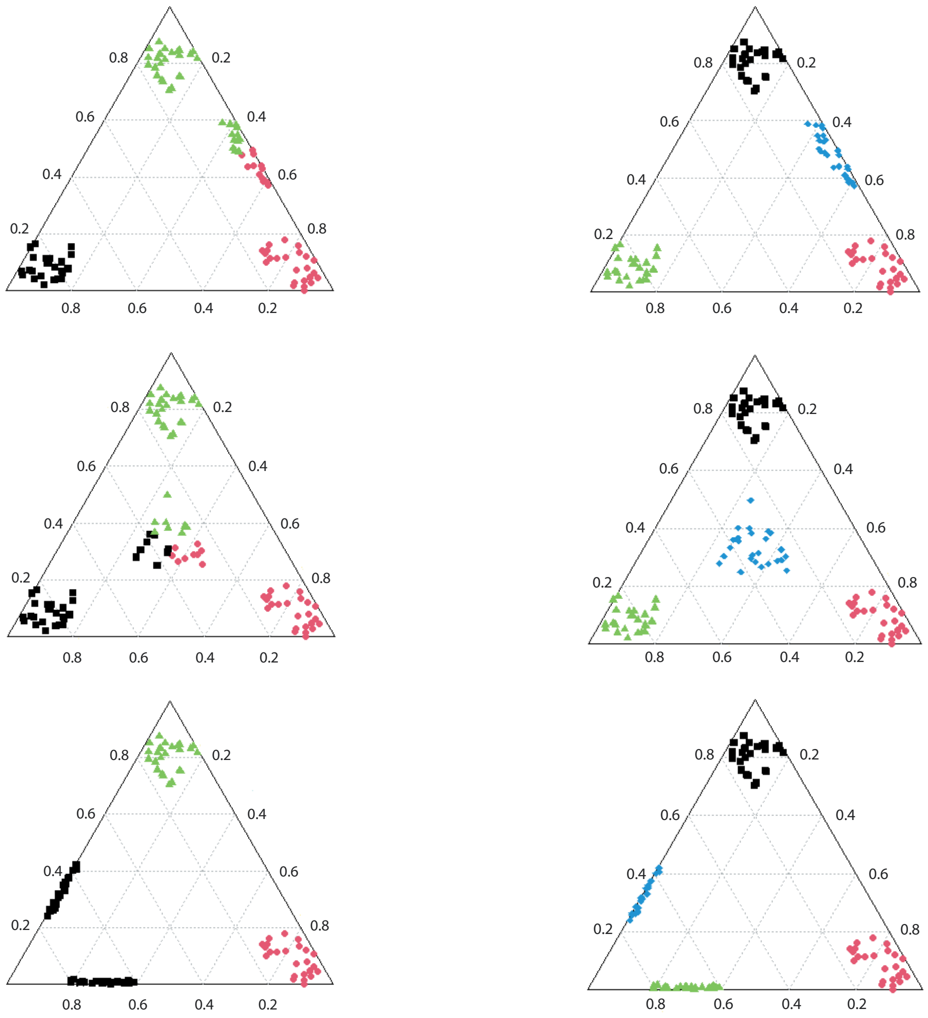
The Hellinger distance can be defined for a general random variable in terms of a measure theory and simplified for the discrete and continuous random variables. In particular, let us consider two discrete random variables with a finite number of possible values which occur with probabilities p_i and q_i respectively. If we denote H as the Hellinger distance between these two discrete probability distributions, then $1 - \Sigma(p_i\, q_i)^{1/2}$ is equal to its second power H^2.

Application of the Hellinger distance to the results of archetypal analysis

Recall that the output of the archetypal analysis consists of the archetypes and the coefficients which approximate the observations as convex combinations of the archetypes. It means that they are weighted averages of the archetypes with nonnegative weights, which have a sum equal to one, and therefore can be interpreted as probability distributions. They indicate probabilities for the given observation to belong to the classes represented by the archetypes (Cutler & Breiman, 1994).

Our proposed methodology consists of identifying the observations with these probability distributions. Then, the distance matrix, needed as an input to hierarchical clustering algorithms, is the matrix of Hellinger distances between the probability distributions. In Figure 1 we show that this approach recovers clusters which are intuitively expected in several scenarios and compares it with a simple approach which assigns the observation to the archetype with the highest weight. For a better graphical representation of the examples, we use probability distributions with three possible values (corresponding to three archetypes). It allows us to plot them using ternary graphs, for which we use the vcd package (Zeileis, Meyer & Hornik, 2007) for the R software.

Figure 1. Clustering of the simulated data of probability distributions with three possible values, visualized by a ternary plot



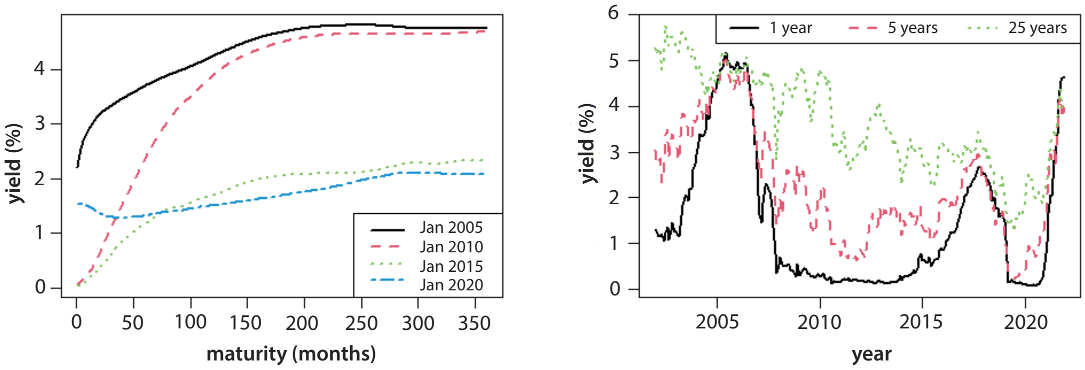
Left: Clusters are determined by the highest weight among the archetype weights. Right: The proposed methodology with the number of clusters chosen by silhouette criterion (Rousseeuw, 1987) implemented in a *cluster* R package (Maechler, Rousseeuw, Struyf, Huber & Hornik, 2019) and hierarchical clustering with single, complete, average and Ward's method. All the methods lead to the same clusters for these data.

APPLICATION TO YIELD CURVE DATA

Data

We use monthly observations for the zero-coupon yield curve in the USA during the period 2003-2022, *i.e.*, 20 years. For every month, the interest rates with 360 maturities are available from 1 month to 30 years: 1 month, 2 months, …, 360 months. The methodology of the yield curve construction is given in (Liu & Wu, 2022) and the data are available on the author's website (Liu & Wu, 2023). We show a selection of the dataset in Figure 2.

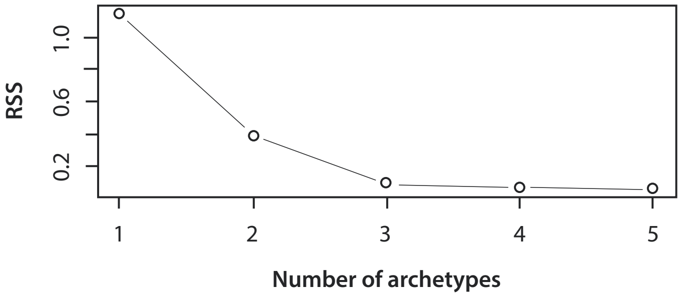
Figure 2. Sample of the data. Left: selected yield curves, right: time evolution of the yields for selected maturities



Archetypal analysis

The elbow criterion applied to residual sums of squares, as suggested by (Eugster & Leisch, 2009), determines the optimal number of archetypes to be 3, see Figure 3. Since the constrained optimization in the archetypal analysis uses a penalization approach, the equality constraint on the sum of the weights is not satisfied exactly. Instead of the required value 1, it ranges between 0.9905 and 0.9989. Before the subsequent analysis, we scale the weights for each of the observations by dividing them by their sum.

Figure 3. Selection of the number of archetypes



We show the yield curves corresponding to the archetypes in Figure 4. We can see that two of the archetypes are yield curves that correspond to high and low values of the interest rates in the dataset. The third archetype is an increasing yield curve. In Figure 5 we show the time evolution of the weights, corresponding to each of the archetypes. Clearly, in the different time periods, different archetypes are dominant.

Figure 4. Archetypes for the yield curves

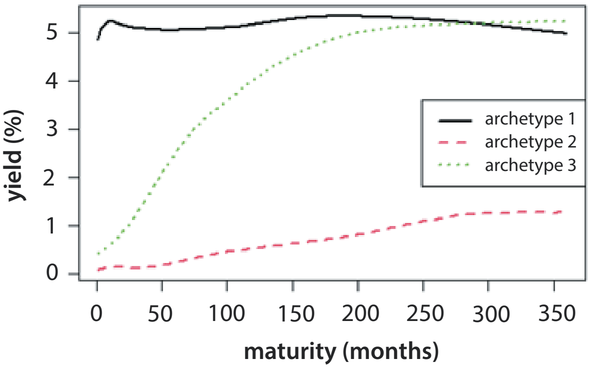
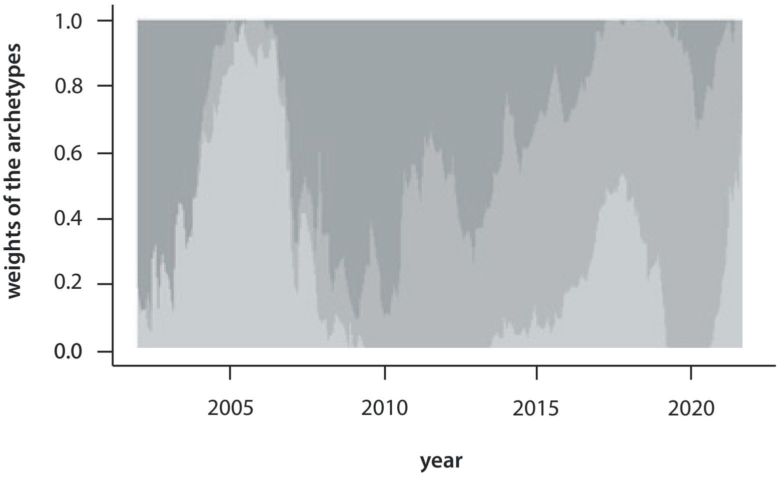


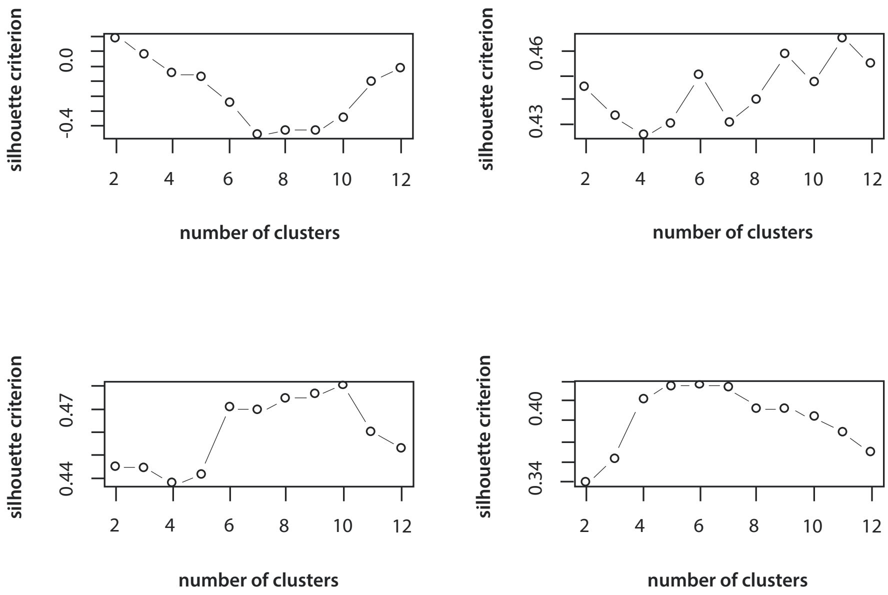
Figure 5. Time evolution of the weights: archetype 1: light grey, archetype 2: grey, archetype 3: dark grey



Clustering based on the Hellinger distance

As described in the methodology section, we identify the yield curves in the given month with a triplet of weights, corresponding to the archetypes. We consider these triplets as probability distributions and compute a matrix of Hellinger distances between them. These distances are then input to hierarchical clustering. We consider single, average, Ward and complete linkage and for different choices for the number of clusters we evaluate the value of silhouette criterion (Rousseeuw, 1987), a simple measure of cluster separation which has become very popular, *cf.* (Wierzchoń & Kłopotek, 2018). We use the implementation by Maechleter *et al.* (2019). The results are shown in Figure 6. The plot for the single and average linkage methods does not show a clear maximum. The Ward linkage leads to a relatively high number of clusters. We therefore consider the complete linkage. Since the value of the silhouette criterion for 5 and 7 clusters is very close to the optimal, we will compare the corresponding clusters with the results obtained by the selected optimal 6 clusters.

Figure 6. Selection of the number of archetypes. The silhouette criterion is applied to the results of hierarchical clustering using single (top left), average (top right), Ward (bottom left) and complete (bottom right) linkage



As in the illustrative examples in the methodology section, we again have three archetypes and therefore we can plot the weights using ternary plots. In Figure 7 we show the weights, distinguishing them according to the clusters obtained for 6 clusters and Figure 8 shows all the yield curves from the dataset and the curves from the individual clusters. This allows us to see that the clusters indeed contain visually similar curves, as well as their difference from the remaining observations. Time periods, in which the yield curves fall into each of the clusters are presented in Table 1. As we can see, this segmentation of the 20 years of data is a reasonable trade-off, when trying to achieve two goals: changing the clusters when the yield curve changes its shape and avoiding too frequent changes which would prevent us from identifying time periods with a certain regime.

Figure 7. Clustering the weights into 6 clusters using Hellinger distance and complete linkage

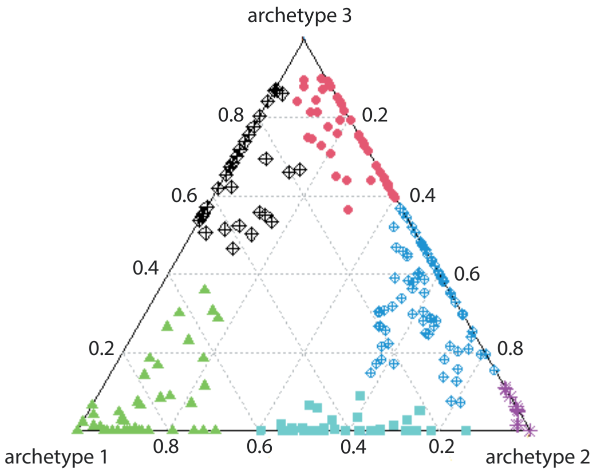


Figure 8. Grey: all yield curves in the dataset, black: yield curves from clusters 1-6 as defined by Table 1 (by rows)

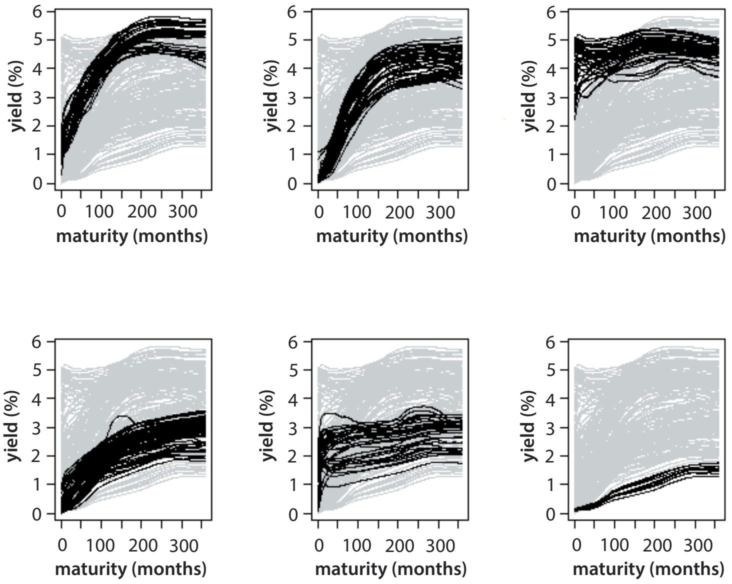


Table 1. Segmentation of the time period of the dataset (2003-2022)

Cluster	Months
1	January 2003 - April 2003 July 2003 - December 2004 January 2008 - October 2008
2	May 2003 - June 2003 November 2008 January 2009 - August 2011 June 2013 - April 2014
3	January 2005 - December 2007 Septemer 2022 - December 2022
4	December 2008 September 2011 - May 2013 May 2014 - October 2017 January 2021 - December 2021
5	November 2017 - February 2020 January 2022 - August 2022
6	March 2020 - December 2020

Finally, we compare the selected clustering with 6 clusters with those obtained for 5 and 7 clusters. Recall that in hierarchical clustering, each step consists of joining the two most similar clusters. In the case of our data, when going from 6 to 5 clusters, the clusters 4 and 6 are joined. On the other hand, the selected clustering with 6 clusters is obtained by joining two clusters in the last step, which now together form the current cluster 3. One of these clusters contained subperiods of January 2005 - November 2005 and August 2007 - December 2007, while the other one contained subperiods of December 2005 - July 2007 and September 2022 - December 2022.

CONCLUSION

We proposed a clustering technique which connects the archetypal analysis with the Hellinger distance between probability distributions. We applied this method to the data about zero-coupon yield curves in the USA during the 20-year period. Since the observations are ordered by time, the clustering also provides a time segmentation. The results were reasonable from both points of view – the similarity of yield curves in the clusters and the length of the time periods. Also, our hypothesis about the relation to classical yield curve factors – level, slope and curvature – was confirmed. The first archetype is a flat curve with almost zero slope and curvature and its level is close to the highest observed values. The second archetype has a positive slope, but not curvature. Its level is on the other side of the spectrum, corresponding to low-interest rates. Therefore, a combination of these two archetypes produces yield curves with different levels and slopes (we note that not in arbitrary combinations, but their possible values correspond to data, on which the archetypes were constructed), but low curvatures. Curvature is added by means of the third archetype.

The continuation of this research lies in several areas. Firstly, the results, coming from a statistical analysis of the data, need a financial interpretation and looking for factors that lead to similar yields in the obtained time periods. Secondly, it makes sense to look at the yields not as a set of discrete points, but as a curve. This suggests using functional archetypal analysis. However, we also note that the proposed clustering in the current form produces meaningful results and therefore it can be applied to different data sets of various kinds.

ACKNOWLEDGEMENT

The work was supported by VEGA 1/0760/22 Modelling negative interest rates grant.

REFERENCES

Bauckhage, C., & Thurau, C. (2009, September). Making archetypal analysis practical. In *Joint Pattern Recognition Symposium* (pp. 272-281). Berlin, Heidelberg: Springer Berlin Heidelberg.

Canhasi, E., & Kononenko, I. (2016). Weighted hierarchical archetypal analysis for multi-document summarization. *Computer Speech* & *Language, 37*, 24-46.

Cutler, A., & Breiman, L. (1994). Archetypal analysis. *Technometrics, 36*(4), 338-347.

Epifanio, I. (2016). Functional archetype and archetypoid analysis. *Computational Statistics* & *Data Analysis, 104*, 24-34.

Epifanio, I., Ibánez, M. V., & Simó, A. (2020). Archetypal analysis with missing data: see all samples by looking at a few based on extreme profiles. *The American Statistician, 74*(2), 169-183.

Eugster, M. J. A., & Leisch, F. (2009). From Spider-Man to Hero - Archetypal Analysis in R. *Journal of Statistical Software, 30*(8), 1-23.

Fernández, D., Epifanio, I., & McMillan, L. F. (2021). Archetypal analysis for ordinal data. *Information Sciences, 579*, 281-292.

Hannachi, A., & Trendafilov, N. (2017). Archetypal analysis: Mining weather and climate extremes. *Journal of Climate, 30*(17), 6927-6944.

Hellinger, E. (1909). Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik, 1909*(136), 210-271.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics, 22*(1), 79-86.

Litterman, R. B., & Scheinkman, J. (1991). Common factors affecting bond returns. *The Journal of Fixed Income, 1*(1), 54-61.

Liu, Y., & Wu, J. C. (2021). Reconstructing the yield curve. *Journal of Financial Economics, 142*(3), 1395-1425.

Liu, Y. & Wu, J. C. (2023). *Liu-Wu Yield Data*. https://sites.google.com/view/jingcynthiawu/yield-data

Lourenzutti, R., & Krohling, R. A. (2014). The Hellinger distance in Multicriteria Decision Making: An illustration of the TOPSIS and TODIM methods. *Expert Systems with Applications, 41*(9), 4414-4421.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. & Hornik, K. (2019). *Cluster: Cluster Analysis Basics and Extensions.* R package version 2.1.0. [computer software].

Mokhtari, E. B., Landguth, E. L., Anderson, S., & Stone, E. (2021). Decoding influenza outbreaks in a rural region of the USA with archetypal analysis. *Spatial and spatio-temporal epidemiology, 38*, 100437.

Moliner, J., & Epifanio, I. (2019). Robust multivariate and functional archetypal analysis with application to financial time series analysis. *Physica A: Statistical Mechanics and its Applications, 519*, 195-208.

Ning, L., Jiang, X., & Georgiou, T. (2013). On the geometry of covariance matrices. *IEEE Signal Processing Letters, 20*(8), 787-790.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics, 20*, 53-65.

Seiler, C., & Wohlrabe, K. (2013). Archetypal scientists. *Journal of Informetrics, 7*(2), 345-356.

Seth, S., & Eugster, M. J. (2016). Probabilistic archetypal analysis. *Machine learning, 102,* 85-113.

Taheri, S. M., Mahyar, H., Firouzi, M., Ghalebi K, E., Grosu, R., & Movaghar, A. (2017). HellRank: a Hellinger-based centrality measure for bipartite social networks. *Social Network Analysis and Mining, 7*, 1-16.

Tsanousa, A., Laskaris, N., & Angelis, L. (2015). A novel single-trial methodology for studying brain response variability based on archetypal analysis. *Expert Systems with Applications, 42*(22), 8454-8462.

Vamvakaris, M. D., Pantelous, A. A., & Zuev, K. M. (2018). Time series analysis of S&P 500 index: A horizontal visibility graph approach. *Physica A: Statistical Mechanics and its Applications, 497*, 41-51.

Vinué, G., & Epifanio, I. (2017). Archetypoid analysis for sports analytics. *Data Mining and Knowledge Discovery,* 31, 1643-1677.

Vinué, G., Epifanio, I., & Alemany, S. (2015). Archetypoids: A new approach to define representative archetypal data. *Computational Statistics* & *Data Analysis, 87*, 102-115.

Wierzchoń, S. T., & Kłopotek, M. A. (2018). *Modern algorithms of cluster analysis* (Vol. 34). Springer International Publishing.

Zeileis, A., Meyer, D., & Hornik, K. (2007). Residual-based shadings for visualizing (conditional) independence. *Journal of Computational and Graphical Statistics, 16*(3), 507-525.

KLASTERIZACIJA ZASNOVANA NA ARHETIPSKOJ ANALIZI

Rezime:

Arhetipska analiza je tehnika redukcije dimenzionalnosti, koja se zasniva na pronalaženju malog broja reprezentativnih elemenata, nazvanih arhetipovi. Zapažanja se zatim aproksimiraju konveksnim kombinacijama arhetipova. Stoga se koeficijenti konveksnih kombinacija mogu tumačiti kao verovatnoće diskretnih slučajnih promenljivih. Vrednosti varijabli identifikuju klase, predstavljene arhetipovima, kojima posmatranje pripada. Na osnovu ove interpretacije, predlažemo da se koristi Hellingerova udaljenost između distribucija verovatnoće za merenje udaljenosti između posmatranja u skupu podataka i da se koristi kao ulaz za grupisanje. Ovaj postupak je primenjen na mesečne podatke krive prinosa bez kupona u periodu 2003-2022. Identifikovane su arhetipske krive prinosa i posmatrane krive su grupisane u šest klastera. Pošto su opservacije date za određeni period vremena dobijeni klasteri su takođe omogućili segmentaciju za odgovarajući period posmatranja.

Ključne reči:

arhetipska analiza,
raspodela verovatnoće,
Hellingerova distanca,
grupisanje,
kriva prinosa